CrossMark

Comparative analysis of Rank Aggregation techniques for metasearch using genetic algorithm

Parneet Kaur[1,4] · Manpreet Singh[2] ·
Gurpreet Singh Josan[3]

Abstract Rank Aggregation techniques have found wide applications for metasearch along with other streams such as Sports, Voting System, Stock Markets, and Reduction in Spam. This paper presents the optimization of rank lists for web queries put by the user on different MetaSearch engines. A metaheuristic approach such as Genetic algorithm based rank aggregation technique has been proposed and implemented in MATLAB for Kendall-tau as (GKTu) and Spearman's foot rule as (GSFD) distance measures. A comparative analysis has been carried out between ranked lists for with and without GA on the basis of simulated results. From the results it has been found that proposed GA optimized rank list (for a particular query on the basis of minimum distance) is better than the conventional methods. In addition, a word association technique i.e., AND-OR operator has been applied on each query. The results are investigated in comparison to non- logic operators for the same query.

Keywords Rank aggregation · Kendall Tau · Spearman's foot rule · Genetic algorithm · Metasearch

✉ Parneet Kaur
  kaur.parneet@gmail.com

  Manpreet Singh
  mpreet78@gmail.com

  Gurpreet Singh Josan
  josangurpreet@rediffmail.com

[1] Research Scholar, Department of Computer Science and Engineering, I.K Gujral PTU, Jalandhar, Punjab, India

[2] Department of Computer Science and I.T, GNDEC, Ludhiana, Punjab, India

[3] Department of Computer Science and IT, Punjabi University, Punjab, India

[4] Department of Computer Science and Engineering, Ambala College of Engineering and Applied Research, Devsthali, Ambala, India

1 Introduction

During the last two decades Page Ranking for web queries has gained great interest by the machine learning community. Due to ubiquitous nature of ranking it is also applied to different fields such as MetaSearch, voting method in elections, information retrieval and collaborative computing (Aledo et al. 2013). Rank aggregation techniques are mainly designed for web based tools to provide effective and fast response to search queries put on by the users. In the current scenario, web based activity raised beyond the imagination around the world with new and very advanced MetaSearch engines. So, the demand of fast query searching techniques has been raised every day. There are plenty of search engines available in the worldwide market which differs in terms of their search capabilities, speed and indexing algorithms. Hence, new methods and models are required to improve the performance of search engine in terms of finding query near to or exactly correct as per the user. Another very important factor is response time taken by the search engine to a particular query. Web search evaluation system is the coordination of search based algorithms along with distance evaluating methods (known as Rank Aggregation Methods). Generally, a user will always be interested in viewing first twenty top results for his query which is displayed on desktop screen. But for a particular query a search engine has to evaluate the distance of this query w.r.t to the web pages linked to this engine. Sometimes results are found to be satisfactory on the first page but, in some other cases favourable results may or may not be found. Hence by improving ranking of web pages to a particular query will provide us better and fast search options.

2 Earlier work done

Rank aggregation might look simple but according to Social Choice theory, it is impossible to design a fair voting ranking list. First voting paradox was discovered by a French mathematician M. Condorcet in 1785 (Condorcet 1785) but before that in 1781 J.C Borda compiled his work highlighting the importance of ranking in field of Voting (Borda 1781). The work reported by him described the situation such that there is no winner for an election. Later, in 1951, Kenneth Arrow published the impossibility theorem which defined the best outputs for social choice theory. According to Arrow, it is impossible to design an aggregation method for three or more than three candidates who meet all required criterions of fairness (Arrow 1951). Lot of research have been proposed which highlights the solution for Rank aggregation methods but, none of them is perfect as proved by Arrow in his theorem. The Rank Aggregation technique is divided into two areas i.e., unsupervised learning approach and supervised learning. In case of unsupervised approach, no training dataset is required for deriving results like Borda Count (Borda 1781; Saari 2000), MedianRank aggregation (Dwork et al. 2001), Genetic and Fuzzy based rank aggregation (Beg and Ahmad 2002) whereas, the supervised techniques utilizes a dataset. To achieve efficient dataset in context of web page ranking, the related work can be compared from earlier which was related to Social Choice Theory (Bartholdi et al. 1989) as it reduces the discrepancy among various

similar data fed to search engines as well as generates the merged result. The work done by Diaconis and Graham (1977) first applied the two popular distance measures i.e., Spearman's footrule and Kendall Tau between two lists of Social choice theory and calculated the distances between them. Spearman's footrule distance calculate the sum of absolute difference between the ranks obtained according to the two lists whereas, the method by Kendall (1938) calculates the pair-wise disagreements between two lists and satisfies the neutrality and consistency mentioned under the Condorcet Property (Young 1974, 1988).

Larry Page and Brin (1998) placed the biggest milestone for exploring World Wide Web. The Page Rank algorithm helps in calculating the relevancy of pages in comparison with other existing pages. PageRank (Page and Brin 1998) and Kleinberg's HITS algorithm (Kleinberg 1999) are two important algorithms for measuring the importance of document on web. Also, the work reported by Dwork et al. (2001) mainly focused on the problem of combining ranking results from various sources in context of web. The method is similar to the Page Rank (Page and Brin 1998) graph based method as it also uses Markov Chain concept for generation of new list of web pages. Montague and Aslam (2001) has applied of rank aggregation in metasearch and decomposed the problem into normalization score technique. Montague et al. (2001b) addressed relevance score normalization (Fox et al. 1999; Lee 1997) for not retrieved documents. Renda and Straccia (2003) covered the rank fusion as a problem under the area of MetaSearch in which various lists were combined in such a way that an optimize solution was being produced after combination. Akritidis et al. (2011) have dealt with designing of a MetaSearch engine named as QuadSearch engine which aimed to combine speed, reliable rank aggregation method, "spam" free result along with detailed and enriched information.

The concept of multiple objective GA based rank aggregation techniques is addressed by Kaur et al. (2015). The Mean and Stuart methods have been taken in consideration for rank order related to queries. Comparative analysis using genetic algorithm between Stuart and Mean method with genetic based approach is done. The performance for GA based approach came out to be the best but with more time.

3 Methods for rank aggregation

3.1 Kemeny young method

Kemeny Young method was suggested by John Kemeny (1959). Young and Levenglick (1978) showed that Kemeny method satisfies Condorcet criterion. Condorcet criterion indicates the robustness of a method where a candidate 'c' elected as winner if it beats every other candidate in a pairwise comparison. Kemeny Young method produces the ordering of candidates on the basis of minimized sum of the Kendall tau distance. Unfortunately, Kemeny's approach is found as NP-hard, even when it used for four rank lists. Dwork et al. (2001) has proposed a 2-approximation to Kemeny Young method employed for Spearman's foot-rule instead of Kendall-tau distance.

3.2 Borda count method

Jean Charles de Borda (1781) in late 1770s developed an aggregate rank list of candidates from political elections. Saari (2000) had applied Borda count in which score was assigned to each candidate based on its position in a rank list, e.g., if a candidate 'c' is in first position of a list of size $n \times 1$, then this candidate will assigned with a score of n −1 since it beats n-1 candidates. Similarly, candidate ranked second will have a score of n-2 and so on. Whereas candidate at the bottom of a rank list will be assigned with zero score as it beats no other candidates. Cumulative score of a candidate in the final ranked list was used to rank candidates in a list. The method of Borda comprised of rank R given to the item n and a list $mk \in R$. The $A_{mk}(i)$ is equal to the number of items j in mk such that $mk(j) > mk(i)$. The total Borda score for the item i is give as.

$$A_{mk}(i) = \sum_{m \in R} A_m(i) \qquad (1)$$

Allocation of rank is based on sorting the score value as A_t where, score with large value is given small rank. Equation (1) represent the method for a list m_k, where $|m_k| = |U| - d$, and $A_{rk}(i)$ is calculated for all items, $i \in m_k$, for all items $j \neq m_k$ assigned as.

$$A_{mk}(j) = \frac{(d+1)^2 + (d+1)}{2d} \qquad (2)$$

Borda Count has very simple computations and can be easily implemented. However, Borda is not a Condorcet method and it also fails few other criterions of fairness.

3.3 Markov chain based methods

Dwork et al. (2001) had proposed four Markov chain based methods. These methods were based on a Markov chain which followed transitions between ordering of candidates in the rank lists. The rank of a candidate was derived by stationary distribution of the Markov Chain. In his PageRank algorithm was based on Markov Chain method.

3.4 The mean by variance (MBV) method

The MBV method comprised of function related to Fuzzy based method which derives an aggregated list. The MBV method is related to Cardinal rank aggregation also called as score based method. M.M. Beg (Beg and Ahmad 2003) used the variance computed from Eq. (3). Higher Rank was given to the

document based on least mean variance position. The ratio for all 'M' documents was derived from Eq. 3 and arranged in an ascending order as aggregated list.

$$mbv(i) = \left(\overline{m}_{di} \Big/ \overline{\sigma}_{di}^2 \right) \tag{3}$$

3.5 The PageRank method

Page and Brin (1998) have developed PageRank algorithm at Stanford University. This algorithm has become the back bone of Google search engine and various other search engines. The Page rank methods mainly calculate the relevancy of pages in comparison to other existing web pages. Other parameters like title, tag, anchor tag, keywords are also used to calculate the overall rank of page under PageRank.

Steps used for Page ranking are given under SERPS which are as follows (Page and Brin 1998):

a)   The pages which carry keywords are first filtered out.
b)   Ranking is performed according to maximum keyword occurrence in a page.
c)   The value for inbound anchor text links is calculated.
d)   Finally the results are derived in descending order using Eq. 4.

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \tag{4}$$

Where d is a dampening factor and its value is 0.85. The value of d can varied as per user's requirement and following the direct links and $(1-d)$ as the page rank distribution from non-directly linked pages.

3.6 HITS method

Klienberg's *HITS* algorithm (Kleinberg 1999) used the concept of authority and hub instead of in-links and out-links concept for allotting rank to the pages. The number of edges among hubs and authority were taken for calculation. Weights were assigned to hubs and authority in order to modify the existing results. The algorithm works in the form of iterations for calculation of ranks.

As first iteration start, the authority and hub weight assigned to a node is computed which is given as.

$$a_i = \sum_{j \in B(i)} h_j \ and \ h_j = \sum_{i \in F(j)} a_i \tag{5}$$

The algorithm keeps on iterating until the vectors converge. A graph is generated by *HITS* algorithm by keeping into account the structure of the graph around the node for computation of hub and authority scores.

3.7 Distance measure methods

Distance evaluation among the document list plays an important role for rank aggregation techniques. Basically two distance measures, Kendall-tau distance and Spearman's footrule distance were first reported by Diaconis and Graham (1977) and Diaconis (1988).

a) *Kendall-tau distance measure:* This method is widely used to calculate the correlation between two rank lists. Kendall-tau (Kendall 1938) distance counts the number of inversion in the order of pair of candidates between two rank lists. The number of pair-wise disagreements is counted by the Kendall-tau distance between two full lists. The distance variable taken for two full lists are denoted by σ and τ as given in Eq. 6:

$$K(\sigma, \tau) = |\{(k, l); k < l, \sigma(k)\langle\sigma(l) but \ \tau(k)\rangle\tau(l)\}| \qquad (6)$$

The normalized value for Kendall-tau distance derived by dividing with greatest possible coefficient value of 0.5 (Kendall 1938). Another method is Optimal aggregation method which works on the principle of removal of outliers from various methods of aggregation. This method satisfied the neutrality and consistency called as Condorcet property (Young 1974, 1988).

b) *Spearman's footrule distance measure:* The Spearman's footrule distance is defined as absolute distance between positions of candidates in two ranking lists. Suppose two document lists $l1 : (d1, d2, d3, d4)$ and $l2 : (d4, d3, d2, d1)$ are given. Spearman's footrule distance d is calculated as the sum of absolute difference between positions of documents, e.g., position of $d1$ in $l1$ is 1st and $l2$ is 4th, therefore, the contribution of document $d1$ to distance is $|1-4|$. Hence total distance is calculated below.

$$d = |1-4| + |2-3| + |3-2| + |4-1|$$
$$d = 8$$

The Spearman's footrule distance calculates the sum of all elements i.e., k for a given set 'A' is generated. The Eqs. 1 and 2 gives the absolute difference between the ranks of k obtained according to the two lists. The two given lists are σ, τ and their Spearman's footrule distance is calculated as:

$$F(\sigma, \tau) = \sum_{k} |\sigma(k) - \tau(l)| \qquad (7)$$

4 Rank aggregations as an optimization problem

The rank aggregation based problems are basically divided into two categories as:

> *Ordinal rank aggregation:* This method employs order base preferences which are provided by voters and this order of candidates helps in calculating the aggregated ranking.
> *Cardinal rank aggregation:* This method based on score provided by the voters to each candidate and this score is used to derive aggregated ranking. In this work both methods are used for generating the consensus ranking.

The problem of Rank aggregation is concerned with finding a consensus ranking list that represent combined list from various search engines. This list is as close to all individual rank lists. In order to find such a combined list is a NP-Hard problem because it cannot be assured that the obtained aggregated list is the best one. Hence, some optimizing techniques are required to implement and solve NP-Hard based problem. Dwork (2001) suggested Kendall distance based aggregation as *Kemeny optimal aggregation* which generates the geometric median of the inputs. Dwork et al. (2001) addressed in his work that the Kemeny optimal aggregation was NP Hard. A polynomial time algorithm was constituted for footrule optimal aggregation (PFOA) for full lists. In the last decade, MM Beg and Ahmad (2003) proposed heuristic approach for Partial Footrule optimal aggregation (PFOA) NP-Hard problem by using the Genetic algorithm based soft computing technique. In his work, both positional methods and PFOA were applied to derive optimal solutions. However, results obtained were better with Spearman's footrule distance measure but, not addressed for Kendall-tau. By taking motivation from the literature survey we have applied the Genetic approach for both Kendall-tau and Spearman's footrule distance on a complex data set.

5 Rank aggregations with genetic approach

Genetic approach is also known as metaheuristic approach. Genetic algorithm (GA) approach was firstly, applied by Goldberg (1989) for the NP-hard problem of partial list which is derived from various rank aggregation methods. Multiple objective genetic algorithms were employed using Mean and Stuart methods for calculation of rank order of different web queries by (Kaur et al. 2015). Genetic algorithm differs from traditional search and optimization in many ways. GA has the ability to avoid being trapped in local optimal solution like traditional methods, which search from a single point. Also, GA approach uses the probabilistic selection and work on the chromosomes which are encoded with solutions parameters. The fitness score value is generated by GA is a beneficial way for deriving objective functions. The Kendall-tau and Spearman's footrule distances measures are taken as an objective functions where, j is the aggregated rank and N are the participating partial lists. The $S_i = \{d_1^i, d_2^i, d_3^i, \ldots \ldots, d_n^i\}$ are the set of document based results derived from search engines. M represents the union of participating engines as $U = \cup_{i=1}^{M} S_i.$. Hence union of all set is $U = \{1, 2, \ldots \ldots |U|\}$. In this work, five search engines

are taken i.e., *Google[1], AltaVista[2], Deeperweb[3], Excite[4] and Hotbot[5]*. Generally, no search engine generates full list for a given query hence partial list are derived which need to be converted into full list. The distance measures are uniformly calculated after converting the partial list into full list. Let partial list is t_j and full list is t, with the number of elements in them being $|t_j|$ and $|U|$, respectively. Equations (8) and (9) converts the partial list into full list. The position value for full list t is modified using Eq. (9).

$$t_j\left[t_j\left(i\right)\right] = \begin{cases} unchanged & if\ i \leq |t_j| \\ x|x\ \epsilon U \lor x \notin t_j, & otherwise \end{cases} \tag{8}$$

$$t(i) = \begin{cases} unchanged, & if\ i \leq |t_j|, \\ \dfrac{\sum_{k=\left(|t_j|+1\right)}^{\left(|U|\right)} t(k)}{|U|-|t_j|}, & otherwise, \end{cases} \tag{9}$$

Spearman's footrule distance $F(t, t_1)$ between the lists *t and t_j* is calculated by using Eq. (7) and Kendall-tau distance $K(t, t_1)$ between the lists *t and t_j* is calculated by using Eq. (6). The normalized aggregated footrule distance is computed with the set of N partial document list as $T = \{t_1, t_2, t_3, \ldots \ldots, t_N\}$, by using Eq. (10).

$$F(t, T) = \frac{\sum_{i=1}^{N} F(t, t_i)}{N} \tag{10}$$

Also, the normalized aggregated Kendall-tau distance is given as:

$$K(t, T) = \frac{\sum_{i=1}^{N} K(t, t_i)}{N} \tag{11}$$

The objective functions as given by Eqs. (10) and (11) are required to be minimized using genetic algorithm. Chromosomes for genetic approach are selected as aggregated list t generated from Borda Count, Mean By Variance, PageRank, Markov Chain MC4 and Scaled footrule aggregation SFO methods. The major stages for GA are generation, reproduction, crossover and mutation which are evaluated as per the given NP-Hard problem of partial lists. Firstly, a conversion chromosomes step into decimal form from binary is carried out in order to represent as valid permutation of numbers 1 *to* $|U|$. Then objective function is evaluated with the selected set of chromosomes from the pool of randomly generated permutations of the universe $U = \{1, 2, \ldots, |U|\}$. Also each of the chromosomes from the pool takes its turn to evaluate the objective function, which is the aggregated footrule *distance* and the aggregated *Kendall*

distance along with the rankings given by the *M* participating search engines. Chromosomes with lesser values survived whereas; those having large value are eliminated from the pool.

At the crossover stage, consider two to-be-crossed chromosomes as $D_i^k = [d_{i,1}, d_{i,2}, ..., d_{i,n}]$ and $D_j^k = [d_{j,1}, d_{j,2}, ..., d_{j,n}]$, where $n = |U|$. Exchange of elements within the group of first $m(m < n)$ elements will take place, and remaining elements which are $d_{i,m+1}, d_{i,m+2}, ..., d_{i,n}$ in D_i^k and $d_{j,m+1}, d_{i,j+2}, ..., d_{j,n}$ in D_j^k will get adjusted to form new set of chromosomes which are $[d_{j,1}, ... d_{j,m}, d'_{i,m+1} ....., d'_{i,n}]$ and $D_j^{k+1} = [d_{i,1}, ... d_{i,m}, d'_{j,m+1} ....., d'_{j,n}]$ to form valid permutations. The implementation of permutation will follow the concept of multiplication of permutations . The following chromosomes will be formed after multiplication of permutation as given by Eqs. (12) and (13) as:

$$\left\{ d'_{i,m+1} ....., d'_{i,n} \right\} = \left\{ d_{j,m+1}, d_{i,m+2}, ..., d_{j,n} \right\} \times \left\{ d_{i,m+1}, d_{i,m+2}, ..., d_{i,n} \right\}, \quad (12)$$

$$\left\{ d'_{j,m+1} ....., d'_{j,n} \right\} = \left\{ d_{i,m+1}, d_{i,m+2}, ..., d_{j,n} \right\} \times \left\{ d_{j,m+1}, d_{j,m+2}, ..., d_{j,n} \right\} \quad (13)$$

New generation based valid permutation of the Universe *U* for two distinct chromosomes will be formed as D_i^{k+1} and D_j^{k+1} . The digit to-be- mutated is exchanged with any other randomly selected digit in that very permutation. This process will result in producing chromosomes of valid permutation.

The genetic algorithm for Kendall-tau and Spearman's footrule distance measure is as follows:

Algorithm	Genetic Kendall-tau (GKTu) and Genetic Spearman's footrule (GSFD)

Steps for the proposed Genetic based algorithm as Genetic Kendall-tau (GKTu) and Genetic Spearman's footrule as (GSFD)

1. *Select GA parameters and initial random population // (Chromosomes are taken as population from the pool of randomly generated permutations).*

2. *Execute algorithm and invoke objective functions*

3. *Iteration: set k=1*

   (a) select, crossover and mutate // GA step (Goldberg, 1989)

   (b) compute parameter for distance measures

   (c) compute fmin(distance)

   If (k+1)<n or fmin(distance) >tolerance; Goto 3

4. *Execute GA with new set of parameters as per 3 and compute performance measures.*

5. *end.*

After implementation of GA optimized parameters are obtained corresponds to minimized objective function (Eqs. 10 and 11).

Table 1 Main queries and search engines

Query. No	Query
Q1.	Parallel sorting neural networks
Q2.	Parallel architecture computing
Q3.	Computer information technology

[1] www.google.com

[2] www.altavista.com

[3] www.deeperweb.com

[4] www.hotbot.com

[5] www.excite.com

6 Experiment and results

In this work, three queries are applied as input to five search engines. GA based distance measures such as Kendall-tau (KTu) and Spearman's footrule distance (SFD) are used to aggregate rank lists for these queries. Simulated results show that proposed GA based GKTu and GSFD shows primarily good results as compare to non-GA distance measures. Table 1 shows set of main three queries. The five search engines are taken viz. *Google[1], AltaVista[2], Deeperweb[3], Excite[4] and Hotbot[5]*. The *t*op 20 results from each of these search engines are generated for evaluation. Also, the aggregated results are compared with using Borda, Mean-By-Variance, PageRank, Scaled aggregated footrule and Markov MC4 methods.

Word association technique is one of the novel applications of rank aggregation. All three queries are taken and bind through *AND-OR* operators in order to analyse the effect of word association. The keywords for the given queries are chosen by the user. New set of queries are formulated using *AND-OR* operators separately given in Table 2.

Table 2 Main queries and new logical operator based queries

Query. No	Query
Q1.	Parallel sorting neural networks
Q1a.	Parallel sorting **and** neural networks
Q1b.	Parallel sorting **or** neural networks
Q2.	Parallel architecture computing
Q2a.	Parallel architecture **and** computing
Q2b.	Parallel architecture **or** computing
Q3.	Computer information technology
Q3a.	Computer **and** information technology
Q3b.	Computer **or** information technology

Table 3 Aggregated list derived from Rank Aggregation techniques

Rank Aggregation Method	Aggregated List
Borda Method	t_{query1} = {1 3 2 4 7 9 16 19 12 20 5 10 6 11 15 8 13 14 17 20}
	$t_{query1a}$ = {3 1 4 2 16 5 17 6 12 13 8 11 7 9 15 10 14 18 19 20}
	$t_{query1b}$ = {1 3 5 4 2 18 7 16 14 19 12 6 13 11 8 9 10 15 17 20}
	t_{query2} = {1 2 8 4 6 9 7 10 5 13 3 11 12 14 15 16 17 18 19 20}
	$t_{query2a}$ = {1 2 9 3 5 6 7 15 10 4 8 13 11 12 14 16 17 18 19 20}
	$t_{query2b}$ = {2 1 3 4 5 8 6 9 10 13 7 11 12 15 14 16 17 18 19 20}
	t_{query3} = {11 5 4 2 12 7 8 10 6 9 1 3 19 14 15 16 13 17 18 20}
	$t_{query3a}$ = {8 5 1 2 3 14 10 6 7 16 9 12 11 4 20 15 13 17 18 19}
	$t_{query3b}$ = {2 1 16 3 6 4 12 5 7 10 11 8 9 13 14 18 15 20 17 19}
Mean-By-Variance	t_{query1} = {19 20 16 10 11 9 15 7 4 3 12 5 6 1 2 8 13 14 17 18}
	$t_{query1a}$ = {17 15 9 13 2 6 8 1 12 11 5 16 4 10 3 7 14 18 19 20}
	$t_{query1b}$ = {19 16 18 11 13 14 2 12 6 7 8 9 10 5 3 4 1 15 17 20}
	t_{query2} = {13 11 12 4 7 6 9 10 1 2 8 5 3 15 14 16 17 18 19 20}
	$t_{query2a}$ = {11 15 8 13 7 6 5 9 3 12 10 4 14 1 2 16 17 18 19 20}
	$t_{query2b}$ = {13 4 15 9 7 10 1 6 5 12 14 3 8 11 2 16 17 18 19 20}
	t_{query3} = {4 16 2 11 10 19 12 9 15 1 3 14 6 7 8 13 5 17 18 20}
	$t_{query3a}$ = {1 4 16 20 15 12 14 2 6 17 8 3 7 11 9 13 10 19 5 18}
	$t_{query3b}$ = {16 20 7 11 10 12 8 5 13 4 18 3 9 15 17 1 6 14 19 2}
Markov Chain MC4	t_{query1} = {1 18 2 3 17 14 13 19 4 7 16 11 20 8 9 15 5 6 12 10 }
	$t_{query1a}$ = {1 3 20 2 19 4 18 15 17 5 14 10 16 9 12 13 6 7 8 11 }
	$t_{query1b}$ = {1 3 15 17 20 5 4 8 10 11 2 9 16 19 6 7 8 14 11 13 }
	t_{query2} = {2 1 20 7 19 18 6 8 17 4 9 16 15 5 10 12 14 11 13 3 }
	$t_{query2a}$ = {1 2 20 3 5 19 9 17 18 6 16 14 15 7 12 4 10 11 8 13}
	$t_{query2b}$ = {1 2 20 3 19 4 18 5 17 6 16 8 15 14 7 9 13 10 12 11 }
	t_{query3} = {4 20 5 11 2 18 12 17 19 7 16 8 15 14 6 3 13 1 9 10 }
	$t_{query3a}$ = {8 1 5 19 18 17 2 3 13 14 10 15 20 6 16 9 12 4 7 11 }
	$t_{query3b}$ = {2 1 20 16 19 18 6 17 3 12 15 4 14 5 7 13 11 8 9 10 }
PageRank Method	t_{query1} = {18 17 14 13 11 15 12 8 10 6 5 20 9 16 7 19 4 2 3 1 }
	$t_{query1a}$ = {20 19 18 15 14 10 9 7 8 11 6 13 12 16 5 17 4 3 2 1 }
	$t_{query1b}$ = {15 17 20 16 10 11 9 8 6 14 13 7 12 19 2 18 4 5 3 1 }
	t_{query2} = {20 19 18 17 16 15 14 12 11 3 13 10 5 9 4 8 2 6 7 1 }
	$t_{query2a}$ = {20 19 18 17 16 41 12 11 13 8 4 10 7 15 2 6 9 3 5 1}
	$t_{query2b}$ = {20 19 18 17 16 15 14 13 12 11 10 9 7 8 2 6 5 4 3 1 }
	t_{query3} = {20 19 18 17 16 1 3 15 14 9 13 10 6 8 7 12 2 11 5 4 }
	$t_{query3a}$ = {18 20 17 16 19 4 15 14 12 11 7 9 6 10 13 3 2 1 5 8 }
	$t_{query3b}$ = {20 19 18 17 15 14 13 11 10 9 8 7 5 12 4 3 6 1 2 16 }
Scaled footrule (SFO)	t_{query1} = {19 20 16 6 11 5 12 10 15 4 2 7 9 3 13 14 3 17 18 1}
	$t_{query1a}$ = {17 16 7 9 10 12 13 2 8 6 15 5 1 11 14 4 3 18 19 20 }
	$t_{query1b}$ = {18 19 16 6 8 9 10 14 2 11 13 12 7 15 5 17 3 4 1 20 }
	t_{query2} = {3 13 7 4 9 5 10 12 2 8 11 6 1 15 14 16 17 18 19 20 }
	$t_{query2a}$ = {15 4 9 13 8 7 6 11 3 10 5 12 14 1 2 16 17 18 19 20}
	$t_{query2b}$ = {13 7 4 9 6 10 1 3 5 8 12 14 11 15 2 16 17 18 19 20 }
	t_{query3} = {1 3 11 19 12 4 10 2 6 9 14 15 16 13 8 5 7 17 18 20 }
	$t_{query3a}$ = {4 14 20 8 16 2 1 3 10 6 7 9 12 11 15 13 5 17 18 19 }
	$t_{query3b}$ = {16 12 7 5 4 11 3 8 20 10 9 15 13 18 1 17 2 14 6 19 }

Table 4 Comparison of Kendall-tau Genetic Algorithms (GKTu) with rank aggregation techniques

Query No.	Kendall-tau distance obtained after aggregating the ranks using:			
	Borda's method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.237	0.353	0.266	0.226
Q1a	0.211	0.352	0.295	0.211
Q1b	0.284	0.347	0.314	0.280
Q2	0.100	0.389	0.242	0.305
Q2a	0.105	0.423	0.236	0.370
Q2b	0.047	0.400	0.188	0.380
Q3	0.253	0.373	0.326	0.389
Q3a	0.215	0.357	0.331	0.384
Q3b	0.158	0.326	0.284	0.240
	MBV method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.568	0.363	0.331	0.320
Q1a	0.421	0.321	0.352	0.320
Q1b	0.611	0.352	0.405	0.305
Q2	0.305	0.321	0.331	0.331
Q2a	0.378	0.242	0.315	0.345
Q2b	0.321	0.342	0.367	0.332
Q3	0.400	0.352	0.353	0.342
Q3a	0.447	0.321	0.273	0.331
Q3b	0.563	0.322	0.394	0.341
	Markov MC4 method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.474	0.300	0.363	0.300
Q1a	0.511	0.331	0.378	0.367
Q1b	0.400	0.256	0.350	0.420
Q2	0.500	0.256	0.343	0.233
Q2a	0.468	0.378	0.354	0.268
Q2b	0.436	0.384	0.347	0.350
Q3	0.574	0.386	0.320	0.380
Q3a	0.495	0.400	0.330	0.370
Q3b	0.553	0.356	0.320	0.280
	PageRank method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.758	0.321	0.370	0.320
Q1a	0.795	0.284	0.340	0.330
Q1b	0.732	0.315	0.370	0.300
Q2	0.895	0.321	0.410	0.310
Q2a	0.853	0.342	0.420	0.310
Q2b	0.974	0.284	0.330	0.240
Q3	0.774	0.331	0.310	0.330
Q3a	0.822	0.310	0.370	0.350
Q3b	0.868	0.305	0.410	0.310
	Scaled Footrule method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.600	0.347	0.368	0.336
Q1a	0.478	0.347	0.380	0.370
Q1b	0.600	0.353	0.380	0.330

Table 4 (continued)

Query No.	Kendall-tau distance obtained after aggregating the ranks using:			
	Borda's method	Genetic algorithm with number of generation being:		
Q2	0.226	0.389	0.270	0.360
Q2a	0.342	0.284	0.350	0.360
Q2b	0.236	0.410	0.270	0.340
Q3	0.316	0.384	0.342	0.400
Q3a	0.336	0.430	0.352	0.310
Q3b	0.453	0.330	0.352	0.300

Table 5 Comparison of Spearmen's footrule Genetic Algorithms (GSFD) with rank aggregation techniques

Query No.	Spearmen's footrule distance obtained after aggregating the ranks using:			
	Borda's method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.380	0.520	0.550	0.510
Q1a	0.300	0.500	0.530	0.520
Q1b	0.380	0.450	0.420	0.530
Q2	0.140	0.490	0.410	0.480
Q2a	0.150	0.410	0.530	0.530
Q2b	0.090	0.500	0.450	0.520
Q3	0.310	0.460	0.490	0.570
Q3a	0.320	0.450	0.430	0.530
Q3b	0.240	0.470	0.460	0.520
	MBV method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.730	0.550	0.540	0.420
Q1a	0.530	0.470	0.520	0.400
Q1b	0.740	0.430	0.370	0.400
Q2	0.370	0.500	0.470	0.500
Q2a	0.430	0.550	0.490	0.480
Q2b	0.420	0.440	0.490	0.580
Q3	0.560	0.470	0.470	0.430
Q3a	0.620	0.390	0.390	0.520
Q3b	0.700	0.480	0.480	0.410
	Markov MC4 method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.650	0.460	0.470	0.460
Q1a	0.640	0.510	0.420	0.450
Q1b	0.565	0.395	0.510	0.380
Q2	0.610	0.420	0.510	0.390
Q2a	0.590	0.520	0.510	0.520
Q2b	0.570	0.490	0.500	0.440
Q3	0.730	0.560	0.490	0.500
Q3a	0.670	0.500	0.400	0.470
Q3b	0.720	0.400	0.540	0.560

Table 5 (continued)

Query No.	Spearmen's footrule distance obtained after aggregating the ranks using:			
	Borda's method	Genetic algorithm with number of generation being:		
	PageRank method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.860	0.370	0.300	0.300
Q1a	0.820	0.430	0.570	0.380
Q1b	0.830	0.490	0.400	0.400
Q2	0.990	0.450	0.470	0.400
Q2a	0.960	0.615	0.650	0.535
Q2b	1.000	0.510	0.490	0.440
Q3	0.890	0.520	0.490	0.470
Q3a	0.930	0.510	0.550	0.460
Q3b	0.920	0.560	0.440	0.440
	Scaled Footrule method	Genetic algorithm with number of generation being:		
		10	50	100
Q1	0.725	0.425	0.415	0.435
Q1a	0.600	0.480	0.480	0.450
Q1b	0.690	0.460	0.430	0.480
Q2	0.290	0.490	0.480	0.520
Q2a	0.390	0.470	0.440	0.450
Q2b	0.320	0.480	0.490	0.520
Q3	0.530	0.390	0.490	0.520
Q3a	0.480	0.530	0.340	0.420
Q3b	0.580	0.490	0.470	0.550

6.1 Non GA results for various rank aggregation techniques

After applying all the three Queries of Table 1 to *Google[1], AltaVista[2], Deeperweb[3], Excite[4] and Hotbot[5]* search engines, general lists are obtained. These lists are filtered as full list given in Table 3 by using rank aggregation techniques. Both Kendall-tau and Spearman's footrule distances are calculated for each of the list.

6.2 GA based results for various rank aggregation techniques

The GA based rank aggregation techniques are implemented to obtain optimum value of objective function. The rank aggregation method is segregated on the basis of results having minimum optimized distance. Min distance value will helps to improve the performance of Metasearch. Table 4 represents the comparison of with and without genetic based optimized Kendall-tau distance. The Tables 4 and 5 both show the result for the proposed Genetic based Kendall-tau Algorithm (GKTu) and Genetic based Spearman's footrule (GSFD) for each rank aggregation method with iterations k = 10, k = 50 and k =100.
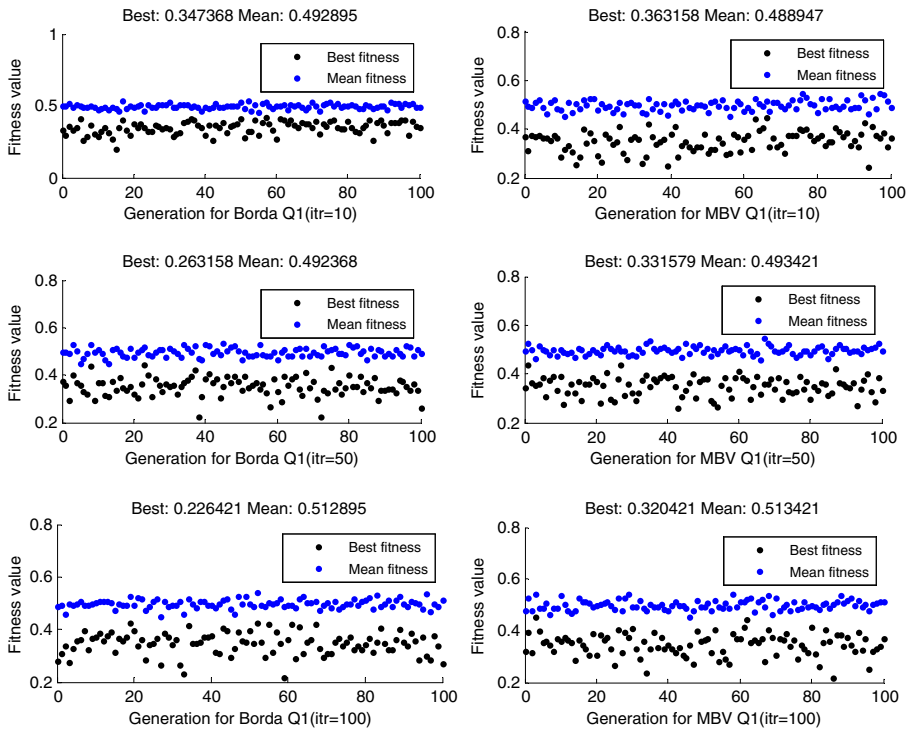
Fig. 1 GA based optimized Kendall-tau distance measure for Borda and MBV Methods

Tables 4 and 5 show the distance values with and without GA, for Borda method. Distance value with improved query 1 by word association technique for Borda with non-GA for Kendall-tau obtained as 0.237 and for the modified as query 1a, the value is minimized to 0.211. Hence, optimized aggregated list are derived for metasearch. Similarly, with GA for Kendall-tau for the same set of queries, value came out as 0.226 and with modification the value is minimized to 0.211. Whereas, in case of *OR* operators the values get slightly increased in comparison to *AND* operator but less than those queries which are without any operator. Hence, the word association based queries helps in improving the quality of search to a larger extent. Figure 1 represents the proposed genetic based Kendall-tau distance measures for Borda and MBV method along with best and mean fitness value for iteration count $k = 10$, 50 and 100. Similarly, Fig. 2 shows the GA based Spearman's footrule distance measure (fitness value) for MC4 and SFO methods with iteration count $k = 10$, 50 and 100.

Tables 6 and 7 reflect the best aggregated list results obtained from the optimized rank aggregation techniques. It has been observed that Borda method is the winner for obtaining the improved results for all types of queries with and without word association operators. Table 8 shows the comparison for all the aggregation methods without GA for both Kendall and Spearman's footrule. Whereas, Table 9 shows the comparison of aggregation methods with GA for both distance measures. Figures 3 and 4 show the
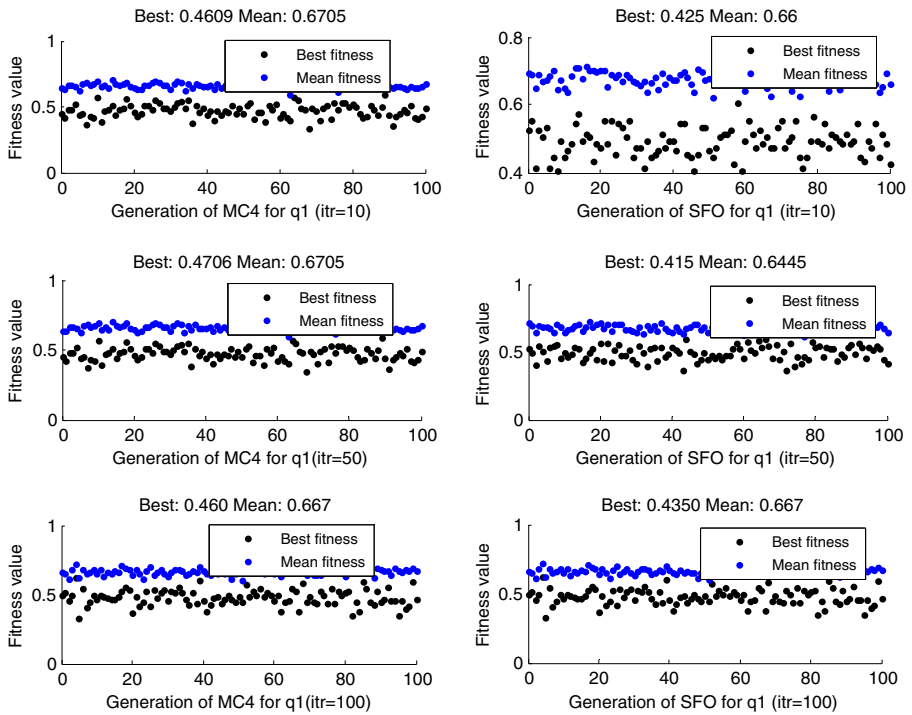
Fig. 2 GA based optimized Spearman's footrule distance measure for MC4 and SFO Methods

Table 6 Top 20 results for the Query Parallel Sorting Neural Network

S No.	URLs for the query Parallel Sorting Neural networks
1.	http://citeseerx.ist.psu.edu/showciting?cid=1148862
2.	http://citeseerx.ist.psu.edu/viewdoc/download?doi
3.	http://link.springer.com/article/10.1007%
4.	http://core.ac.uk/download/pdf/13372685.pdf
5.	http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/cs11/article1.html
6.	http://dl.acm.org/citation.cfm?id=2689060
7.	http://www.personal.psu.edu/lnl/papers/aiaa20057168.pdf
8.	https://en.wikipedia.org/wiki/Artificial_neural_network
9.	http://in.mathworks.com/help/nnet/ug/neural-networks
10.	http://www.neuro.sfc.keio.ac.jp/publications/pdf/hys.pdf
11.	http://dosen.narotama.ac.id/wp-content/uploads/2013/02/7
12.	http://motherboard.vice.com/read/why-googles-neural-networks
13.	http://research.microsoft.com/pubs/175561/asru-2011.pdf
14.	http://www.cs.tau.ac.il/~wolf/deeplearningmeeting
15.	http://www.cs.stir.ac.uk/~lss/NNIntro/InvSlides.html
16.	http://www.gcn.us.es/6BWMC/volume/sn-snps-7.pdf
17.	http://digitalcommons.calpoly.edu/
18.	http://www.eecs.berkeley.edu/~krste/papers/MStrain.pdf
19.	http://www.d.umn.edu/~tkwon/research/papers/KWTA_networks.pdf
20.	http://www.neuro.sfc.keio.ac.jp/publications/pdf/hys.pdf

Table 7 Top 20 results for the Query Parallel Sorting AND Neural Network

S No.	URLs for the query Parallel Sorting AND Neural networks
1.	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1
2.	http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=62417
3.	http://link.springer.com/article/10.1007%2FBF00228719#page-1
4.	http://ieeexplore.ieee.org/Xplore/defdeny.jsp?url=http
5.	https://en.wikipedia.org/wiki/Artificial_neural_network
6.	http://www.researchgate.net/publication/3184115_A_super-parallel
7.	http://www.cs.tau.ac.il/~wolf/deeplearningmeeting/pdfs
8.	http://core.ac.uk/download/pdf/13372685.pdf
9.	http://www.cs.stir.ac.uk/~lss/NNIntro/InvSlides.html
10.	http://www.academia.edu/3046206/A_parallel_neural_networks_algorithm
11.	http://www.ncbi.nlm.nih.gov/pubmed/1338621
12.	http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1
13.	https://books.google.co.in/books?id=ihc
14.	http://dosen.narotama.ac.id/wp-content/uploads
15.	http://www.sigmod.org/publications/dblp/db/journals/pc/pc14.html
16.	http://www.mae.cuhk.edu.hk/~jwang/Publications/Journals/
17.	http://dl.acm.org/citation.cfm?id=1553937
18.	http://www.sciencedirect.com/science/article/pii/0196677482900104
19.	http://www.amazon.com/Parallel-Computing-Springer-International-Eng
20.	http://www.iro.umontreal.ca/~bengioy/cifar/NCAP2014-summerschool

graphical representation of Spearman's footrule and Kendall-tau distances related to all the rank aggregation techniques.

7 Conclusions and future scope

Genetic Algorithm approach is applied for distance measures in order to minimize the distance values for various aggregated lists to generate an aggregated rank. Minimum distance derived from analysis improves the quality of metasearch and helpful for more effective MetaSearch engines. The Genetic based Kendall-tau and Spearman's footrule shows variation in the magnitude for Borda method. In case of proposed GA based approach at 10th generation, there is no improvement seen in the distance values but with the increase in the iterations the values are getting optimized. Similarly for MBV,

Table 8 Performance comparison of different rank aggregation techniques for all six queries without GA

Aggregation Techniques (Without GA)	Aggregated distance[0,1] (Kendall-tau)	Aggregated distance[0,1] (Spearman's footrule)
Borda Method	0.263	0.385
Mean-By –Variance	0.669	0.850
Markov MC4	0.735	0.951
PageRank	1.245	1.366
Scaled footrule Aggregation(SFO)	0.597	0.767

Table 9 Performance comparison of different rank aggregation techniques for all six queries with GA at 100th iteration

Aggregation Techniques (With GA)	Aggregated distance[0,1] (Kendall-tau)	Aggregated distance[0,1] (Spearman's footrule)
Borda Method	0.464	0.785
Mean-By –Variance	0.494	0.622
Markov MC4	0.495	0.695
PageRank	0.466	0.638
Scaled footrule Aggregation(SFO)	0.517	0.724

results for GA are consistently improving whereas for Markov Chain MC4 results improved as we moved from 50th generation to 100th generation. Whereas, in case of non-GA approach distance values derived for MBV, Markov Chain MC4, PageRank and SFO techniques results are not found satisfactory. Results derived after using AND-OR operators are showing significant improvement for both proposed GA and Non GA based measures. These Boolean logics when applied in web based search act as an extremely effective filters for finding the exact information require by the users. It is concluded that GA based approach is found efficient and extraordinary as compared to Non GA based approach in terms of optimized distance. For the future scope of this work, more soft computing techniques need to be explored along with different rank aggregation techniques to minimize the execution time.
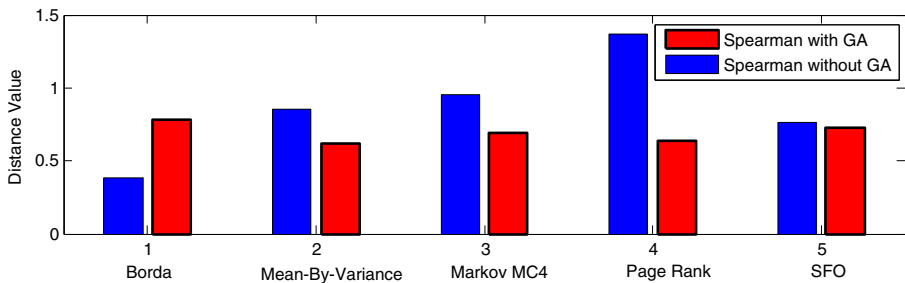

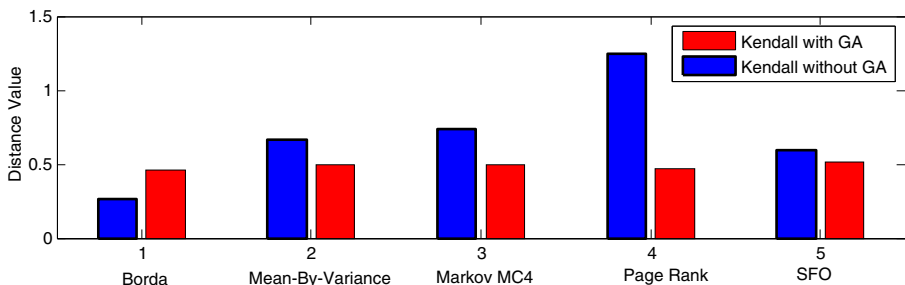**Fig. 3** Distance comparison of rank aggregation techniques for Spearman foot rule

Fig. 4 Distance comparison of rank aggregation techniques for Kendall

References

Akritidis, L., Katsaros, D., & Bozanis, P. (2011). Effective rank aggregation for metasearching. *The Journal of Systems and Software, 84*, 130–143.

Aledo, J. A., Gamez, J. A., & Molina, M. (2013). Tackling the rank aggregation problem with evolutionary algorithms. *The Journal of Applied Mathematics and Computation, 222*, 632–644.

Arrow, K. (1951). *Social choice and individual values*. New York: Wiley.

Bartholdi, J. J., Tovey, C. A., & Trick, M. A. (1989). Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare, 6*(2), 157–165.

Beg, M. M. S., & Ahmad, N. (2002). Fuzzy Logic Based Rank Aggregation Methods for the World Wide Web. In *Proceedings of the International Conference on Artificial Intelligence in Engineering and Technology*, Malaysia (pp. 363–368).

Beg, M. M. S., & Ahmad, N. (2003). Soft computing techniques for rank aggregation on the World Wide Web. *Journal of World Wide Web, 6*(1), 5–22.

Borda, J. C. (1781). Mémoire sur les élections au scrutin. Histoire de l'Académie Royale des Sciences.

Condorcet, M. J. (1785). *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix*.

Diaconis, P. (1988). *Group representation in probability and statistics* (IMS Lecture Series 11). Hayward: Institute of Mathematical Statistics.

Diaconis, P., & Graham, R. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society, Series B, 39*(2), 262–268.

Dwork, C., Kumar, R., Noar, M., & Sivakumar, D. (2001). Rank aggregation methods for the Web. *In Proceedings of ACM Conference on World Wide Web* (pp. 613–622).

Fox, Lee, K., Frieder, O., Knepper, M., & Snowberg, E. (1999). SENTINEL: A multiple engine information retrieval and visualization system. *Journal of the ASIS, 50*(7).

Goldberg, D. E. (1989). Book on Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley.

Kaur, M., Kaur, P., & Singh, M. (2015). Rank Aggregation Using Mutiple Objective Genetic Algorithm. *In the Proceedings of IEEE International Conference on Next Generation Technologies*, Dehradun, India, (pp. 836–840).

Kemeny, K. (1959). Mathematics without numbers. *Daedalus, 88*, 571–591.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1/2), 81–93.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46.

Lee, J. H. (1997). Analyses of multiple evidence combination. *In Proceedings of ACM SIGIR* (pp. 267–276).

Montague, M. H., & Aslam, J. A. (2001). *Relevance score normalization for Metasearch* (In *Proceeding of Conference on Information and Knowledge Management*, (pp. 427–433)). GA: Atlanta.

Montague, M. H., & Aslam, J. A. (2001b). Models of metasearch. *In Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)* (pp. 276–284).

Page, L., & Brin, S. (1998). The anatomy of a large-scale hypertextual Web search engine. *In Proceedings of Seventh International World Wide Web Conference*.

Renda, M. E., & Straccia, U. (2003). Web metasearch: Rank vs. score based rank aggregation methods. *In Proceedings of ACM SAC* (pp. 841–846).

Saari, D. J. (2000).The mathematics of voting: democratic symmetry, Economist. 83. Available at http://vidal.ece.sc.edu/822/papers/econ-voting.html.

Young, H. P. (1974). An axiomatization of Borda's rule. *Journal of Economic Theory* 9, (pp.43-52).

Young, H. P. (1988). Condorcet's theory of voting. *American Political Science Review, 82*, 1231–1244.

Young, H. P., & Levenglick, A. (1978). A consistent extension of Condorcet's election principle. *SIAM Journal on Applied Mathematics, 35*(2), 285–300.